Temporal Aggregation of Visual Features for Large-Scale Image-to-Video Retrieval

Noa Garcia
Aston University

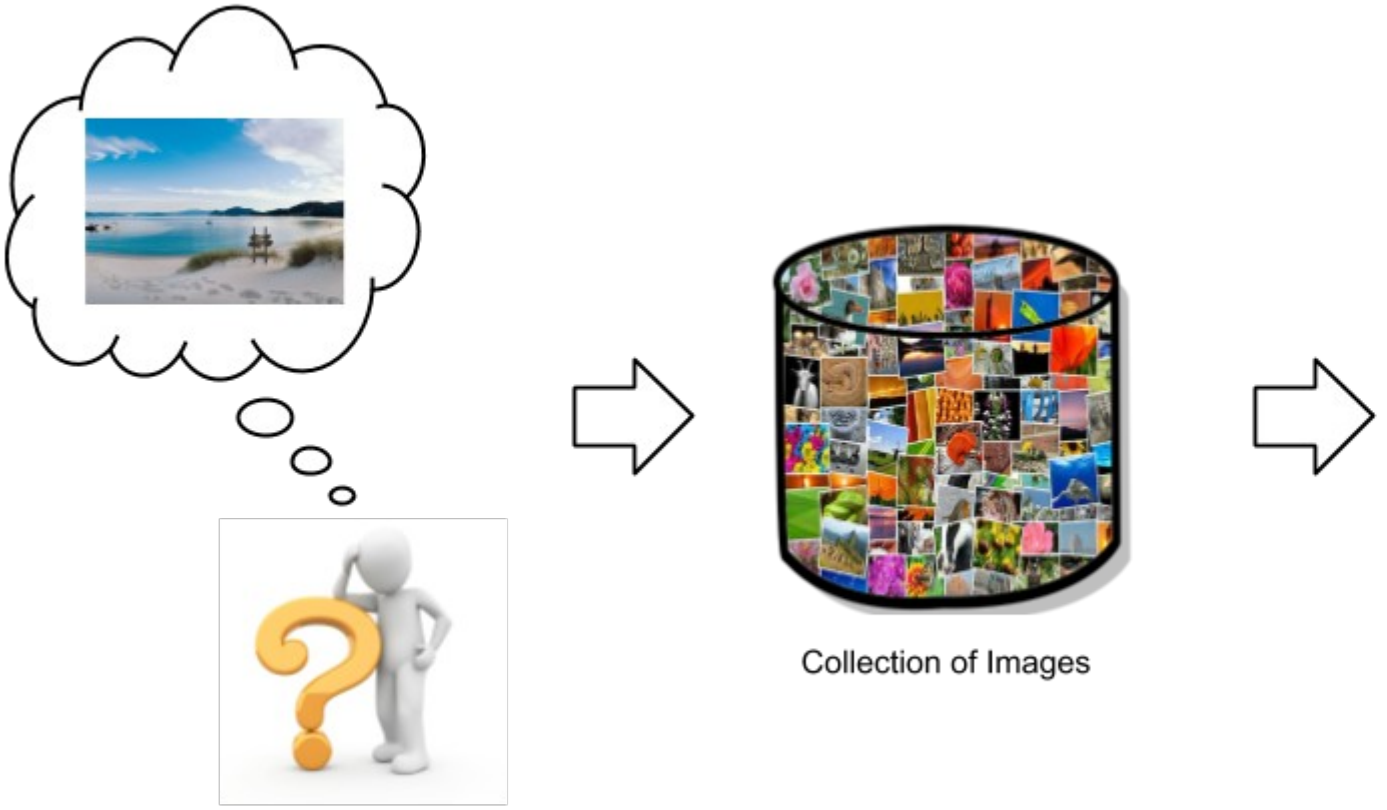
ACM International Conference on Multimedia Retrieval 2018

Motivation

Visual Content Era

Visual Search

- Types of visual search:
  - Content-based image retrieval

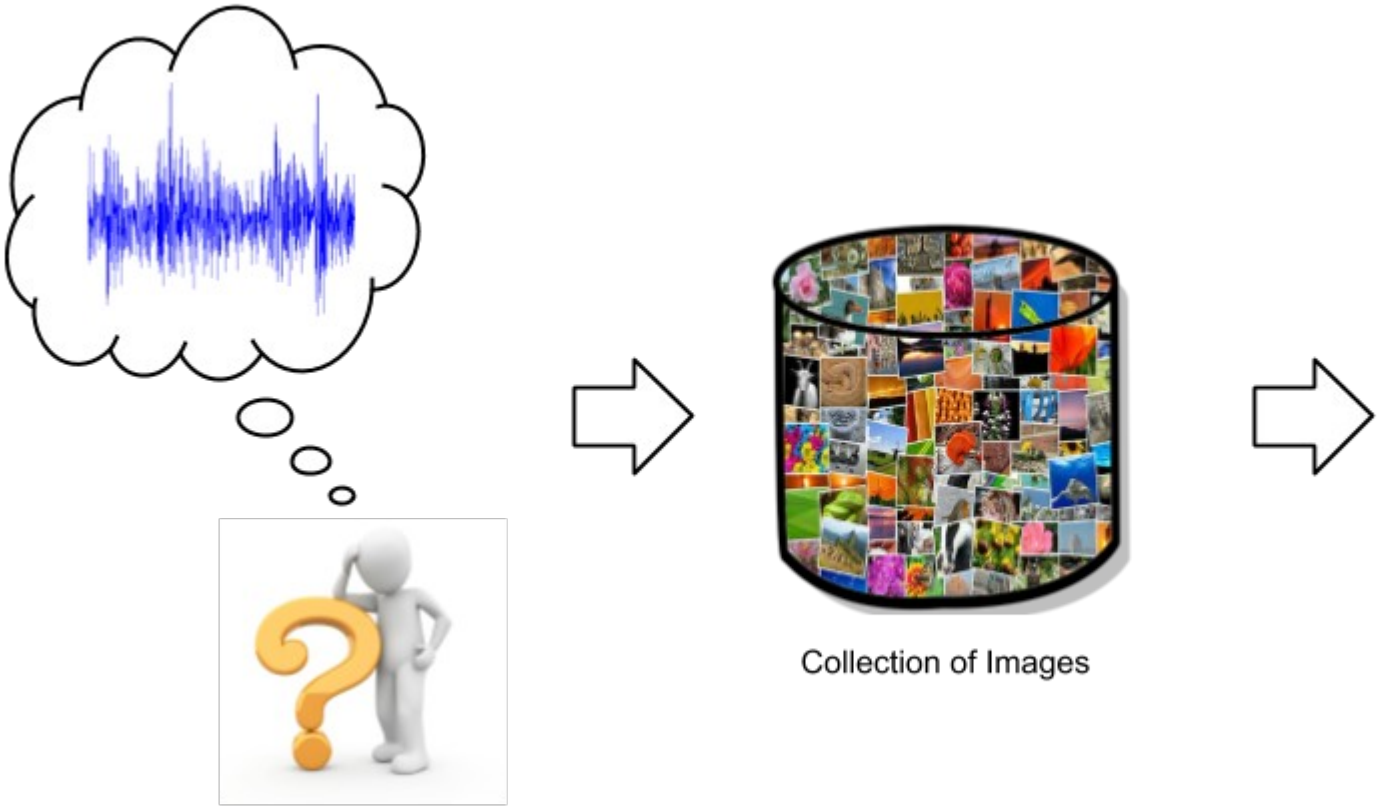


Visual Search

- Types of visual search:
  - Text-to-Image retrieval

Visual Search

- Types of visual search:
  - Audio-to-Image retrieval

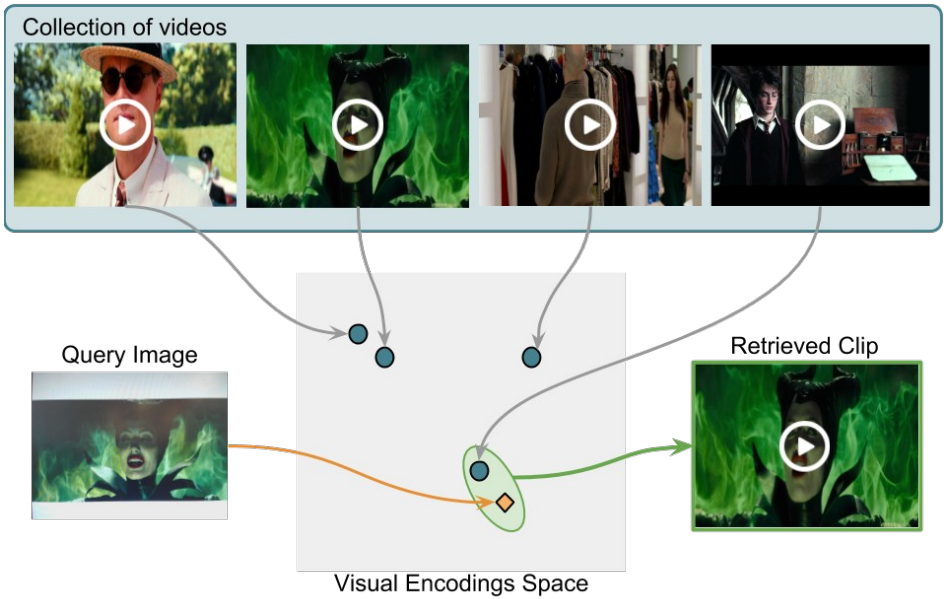


Visual Search

- Types of visual search:
  - Image-to-Video Retrieval

Image-to-Video Retrieval

- Challenges:
  - Asymmetry
  - Temporal Redundancy
  - Scalability
- Objectives:
  - To encode videos and images into a common space
  - To compress redundant data without a dramatic loss in accuracy



Related Work

Related Work

- Early Work: Image Retrieval techniques
  - BoW on frames (Sivic and Zisserman, ICCV'03)
  - Vocabulary Trees on frames (Nister and Stewenius, CVPR'06)

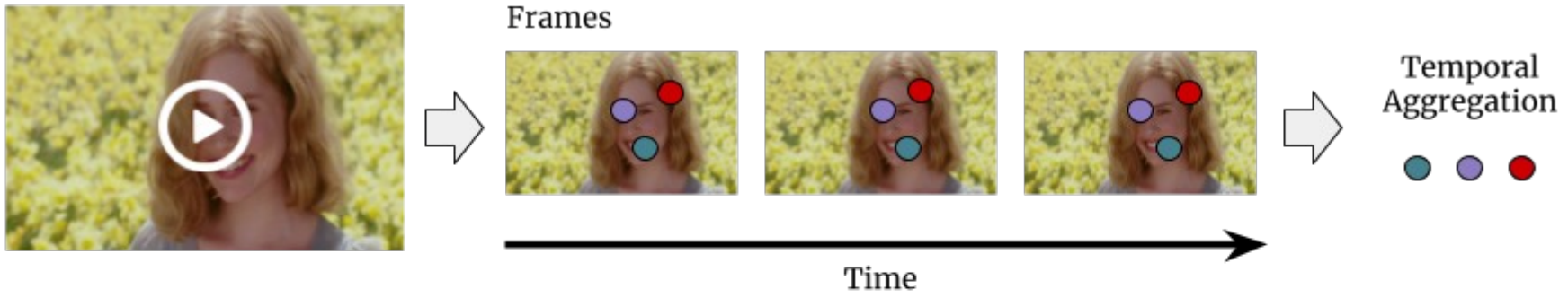
Related Work

- Early Work: Image Retrieval techniques
  - BoW on frames (Sivic and Zisserman, ICCV'03)
  - Vocabulary Trees on frames (Nister and Stewenius, CVPR'06)

- Temporal Aggregation (TA) Methods
  - Based on Local Features
  - Based on Global Feautres

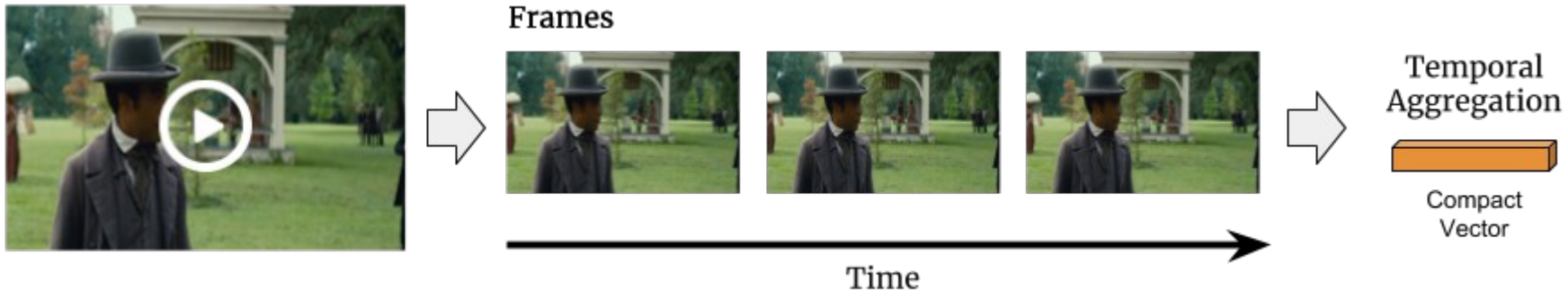
TA: Local Features

- Local features (e.g. SIFT) extracted from each frame and tracked along time
- Tracks are aggregated into a single vector by:
  - Average (Anjulan and Canagarajah, SPIC'07)
  - Minimum distance (Araujo et al. ICIP'14)



TA: Global Features

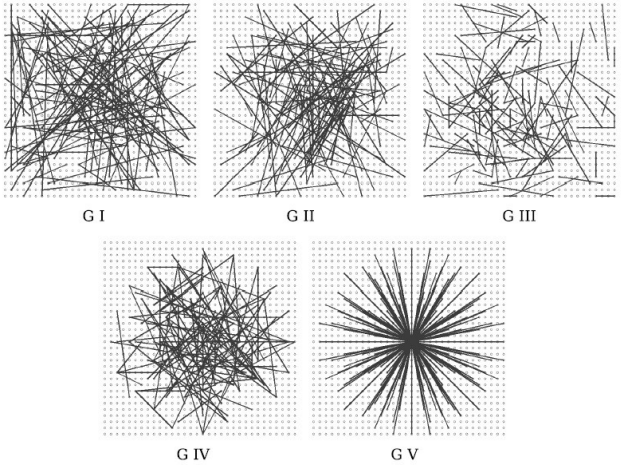
- Encode the visual information of a video segment into a single vector
  - BoW (Zhu and Satoh, ICMR'12)
  - Fisher Vector (Araujo et al., ICIP'15)
  - Bloom Filters (Araujo and Girod, CSVT'17)



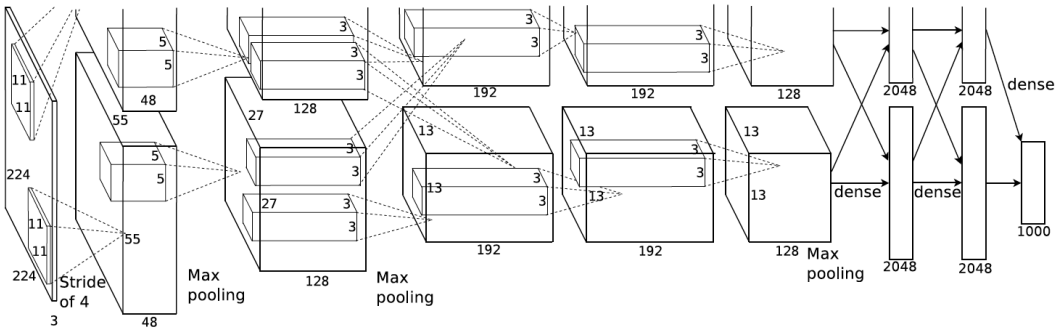
Methodology

Temporal Aggregation

- We propose two models to aggregate temporal information in videos:
  - Local Binary Temporal Tracking (LBTT)
  - Deep Features Temporal Aggregation (DFTA)



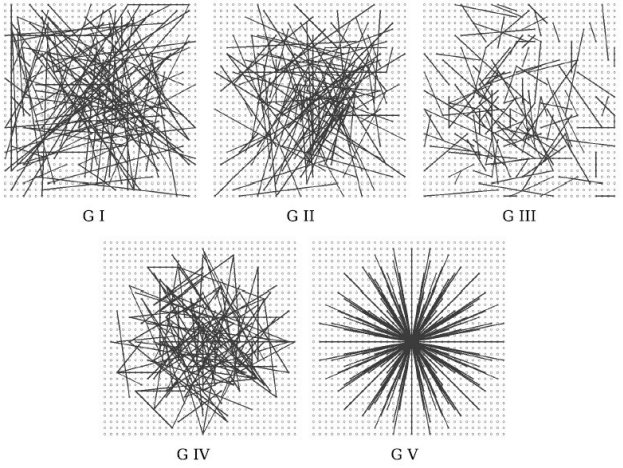
BRIEF descriptor sampling patterns
(Calonder et al., ECCV'10)



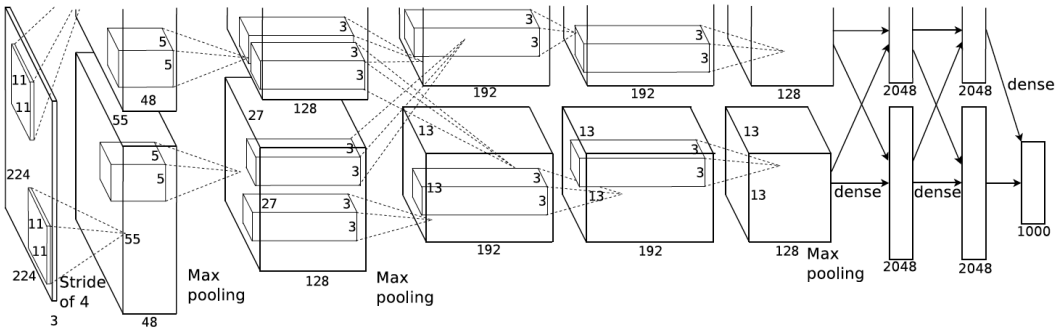
AlexNet architecture (Krizhevsky et al., NIPS'12)

Temporal Aggregation

- We propose two models to aggregate temporal information in videos:
  - Local Binary Temporal Tracking (LBTT)
  - Deep Features Temporal Aggregation (DFTA)



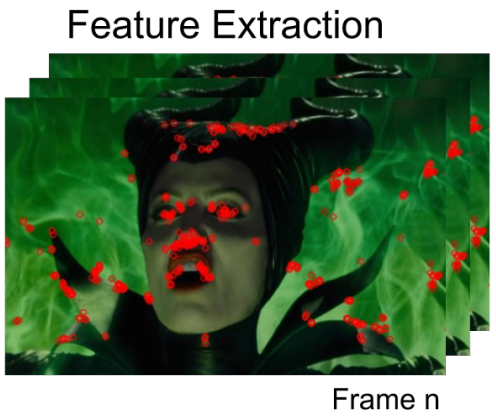
BRIEF descriptor sampling patterns
(Calonder et al., ECCV'10)



AlexNet architecture (Krizhevsky et al., NIPS'12)

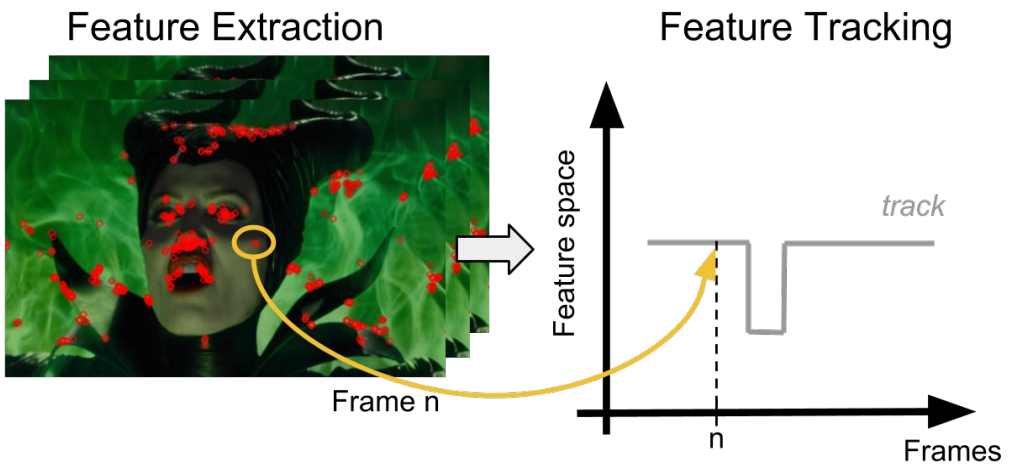
Temporal Aggregation: LBTT

- BRIEF features are extracted from every frame



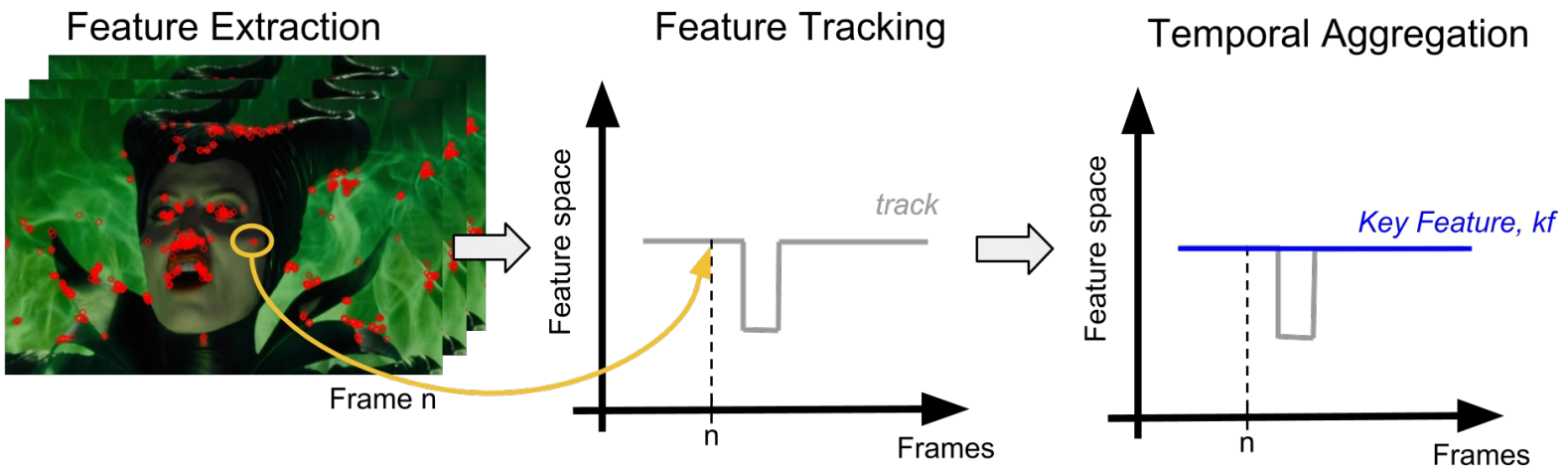
Temporal Aggregation: LBTT

- BRIEF features are extracted from every frame
- Hamming distance to track features along time



Temporal Aggregation: LBTT

- BRIEF features are extracted from every frame
- Hamming distance to track features along time
- For each track, a key feature is computed by majority

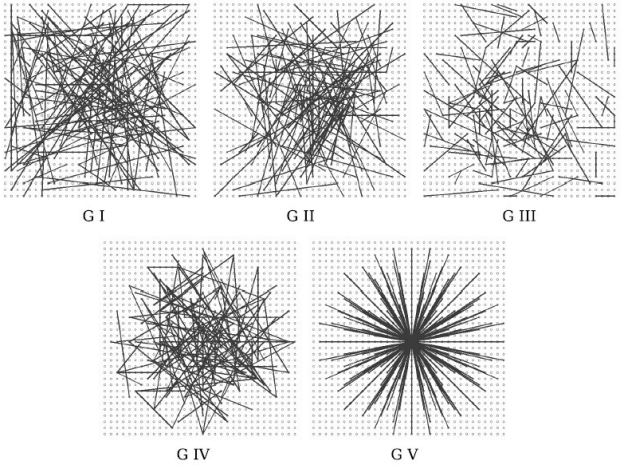


Temporal Aggregation: LBTT

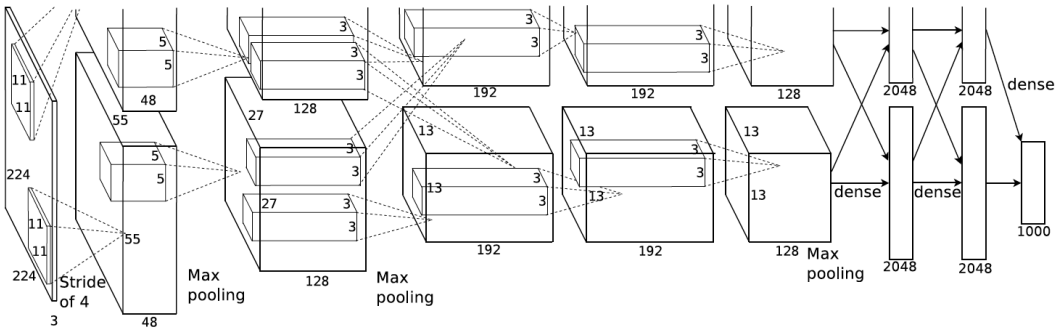
- Shot Boundary Detection
  - When consecutive frames have no common tracks
- Test time:
  - BRIEF features extracted from query
  - Search nearest key features with a kd-tree
  - Key features vote for the shot they belong to

Temporal Aggregation

- We propose two models to aggregate temporal information in videos:
  - Local Binary Temporal Tracking (LBTT)
  - Deep Features Temporal Aggregation (DFTA)



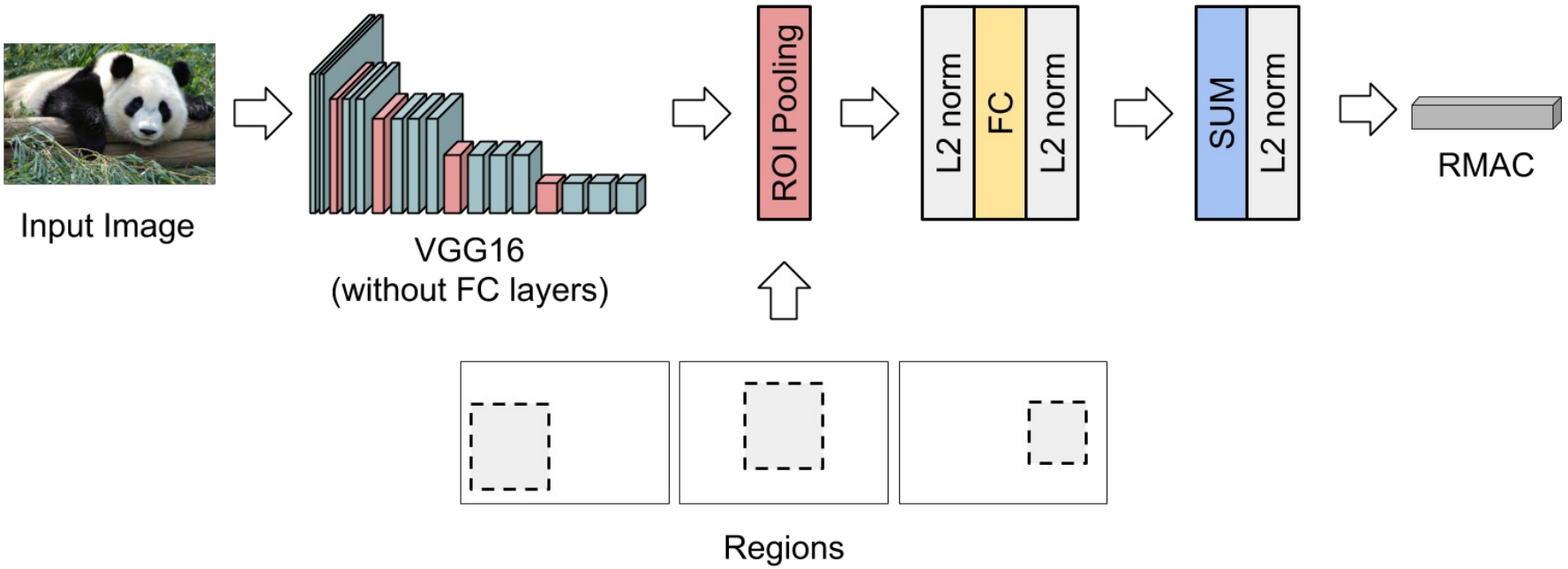
BRIEF descriptor sampling patterns
(Calonder et al., ECCV'10)



AlexNet architecture (Krizhevsky et al., NIPS'12)

Temporal Aggregation: DFTA

- Each frame is encoded using RMAC (Tolias et al., ICLR'16)



Temporal Aggregation: DFTA

- RMACs within the same shot are aggregated by:
  - DLTA-Max: for each dimension, keep the maximum value

$$\Theta(\mathbf{S}_{i,j}) = maxpool(\phi(f_{i,j,k}))$$

  - DLTA-Mean: for each dimension, the average value

$$\Theta(\mathbf{S}_{i,j}) = \frac{1}{N_{S_{i,j}}} \sum_{k=1}^{N_{S_{i,j}}} \phi(f_{i,j,k})$$

Experiments

Dataset

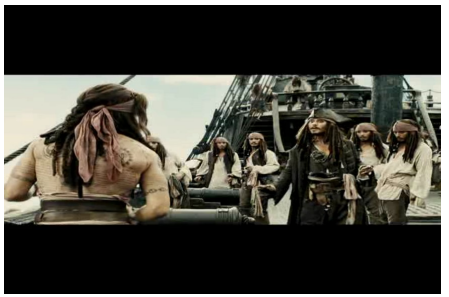
- **MoviesDB** to evaluate image-to-video retrieval
  - Full movies with annotated query images
  - Query images are capture with a webcam
  - Performance measured as accuracy

$$\text{Acc} = \frac{\text{No. Visual Matches}}{\text{Total No. Queries}}$$

The Devil Wears Prada

Groundhog Day

Her



Pirates of the Caribbean: At World's End

Results

- LBTT
  - superior accuracy
  - multiple searches
- DLTA
  - best compression
  - single search per query

Table 1: Results in *The Devil Wears Prada* from MovieDB.

	Method	Dim	Memory	N.Features	Acc
Local	IR-BRIEF	256	2.53 GB	85M	0.93
	LBTT	256	61 MB	2M	0.93
Global	IR-FC1	4096	614 MB	39,324	0.63
	IR-FC2	4096	614 MB	39,324	0.42
	IR-RMAC	512	76.8 MB	39,324	0.91
	DLTA-Max	512	3.13 MB	1,602	0.22
	DLTA-Mean	512	3.13 MB	1,602	0.69

Results

- Results are consistent over different movies
  - LBTT outperforms DLTA in accuracy
  - DLTA-Max poor results

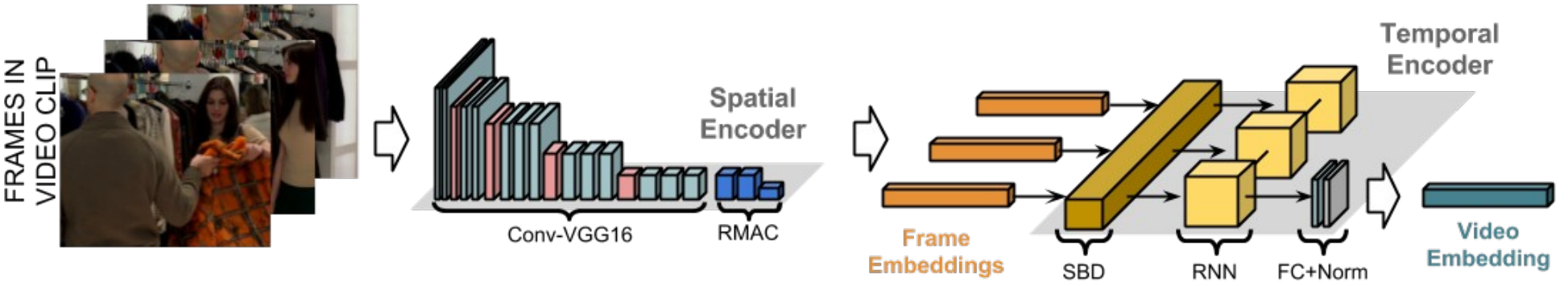
Table 2: Accuracy in The Devil Wears Prada, Groundhog Day, Her and Pirates of the Caribbean movies from MoviesDB.

Method	Movie1	Movie2	Movie3	Movie4
LBTT	0.93	0.97	0.76	0.80
DLTA-Max	0.22	0.16	0.18	0.12
DLTA-Mean	0.69	0.56	0.53	0.47

Future Work

Future Work

- LSTM as Temporal Aggregation Method



- Experiments over more video collections
  - Full MoviesDB dataset (40 movies)
  - Standford I2V (Araujo et al., MMSys'15)

Conclusions

Conclusions

We propose **two temporal aggregation models** for image-to-video retrieval

- A model based on local binary features (LBTT)
- A model based on global deep features (DLTA)

Conclusions

Models based on binary features outperform deep learning models in terms of **accuracy**

Conclusions

However, deep learning models are more **efficient**

Conclusions

Future Work: train better deep learning models to increase accuracy

Thank you!

Noa Garcia
Aston University

Contact: garciadn@aston.ac.uk

ACM International Conference on Multimedia Retrieval 2018

References

- Araujo et al. Stanford I2V: a News Video Dataset for Query-by-Image Experiments. MMSys'15
- Araujo et al. Temporal Aggregation for Large-Scale Query-by-Image Video Retrieval. ICIP'15
- Araujo and Girod. Large-Scale Video Retrieval Using Image Queries. CSVT'17
- Calonder et al., BRIEF: Binary Robust Independent Elementary Features. ECCV'10
- Krizhevsky et al., ImageNet Classification with Deep Convolutional Neural Networks. NIPS'12
- Nister and Stewenius. Scalable Recognition with a Vocabulary Tree. CVPR'06
- Sivic and Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. ICCV'03
- Tolias et al., Particular Object Retrieval with Integral Max-Pooling of CNN Activations. ICLR'16
- Zhu and Satoh. Large Vocabulary Quantization for Searching Instances from Videos. ICMR'12